Machine Learning for EUR-Lex Legal Measures

Ashwin Ittoo, Uliège, BE

19th June 2019
ECB
Frankfurt

EFMD EQUIS ACCREDITED
EFMD EPAS ACCREDITED
AACSB ACCREDITED
HEC LIÈGE Management School - Liège Université

About Myself

- HEC Liège, University of Liège, BE

- Japan Advanced Institute of Science & Tech, JP

- Expertise
  - Machine Learning/Deep Learning & NLP
  - AI & Law
    - *Competition Law: Algorithmic Collusion*
    - *Penal, Criminal Law: Bias & Fairness*

Agenda

- Part 1: Laying out the foundations
  - Scientific, Technical Background
- Part 2: Actual project
  - Machine Learning for EUR-Lex Legal Measures

Part 1: Laying out the Foundations

Machine Learning Intro

- Machine Learning (ML) subfield of AI
- Other AI subfields
  - Robotics
  - Control & Automation
  - Planning & Scheduling
  - Heuristic Search & Optimization
  - …

ML Intro (cont)

- Core principle
  - Train a “machine”
  - Learn how to perform a task
  - From experience collected in the past
- Machine
  - Computer, software, piece of hardware
- Tasks
  - Predict sentiment of customer reviews
  - Predict recidivism risk of offenders
  - Recognize objects in images
- Experience
  - Data collected in the past

How to Learn?

- Main paradigms
  - Supervised Learning
  - Unsupervised (self-supervised learning)
  - Reinforcement Learning
    - Most for agent-based systems
    - Algorithmic Collusion

Supervised Learning (SL)

- Most popular ML paradigm
- Learning from past data…but
- Annotated/labeled with information of interest
  - Sentiment (POS, NEG)

Review:

films adapted from comic books have had plenty of success , whether they're about superheroes (batman , superman , spawn) , or geared toward kids (casper) or the arthouse crowd (ghost world) , but there's never really been a comic book like from hell before . for starters , it was created by alan moore (and eddie campbell) , who brought the medium to a whole new level in the mid '80s with a 12-part series called the watchmen . to say moore and campbell thoroughly researched the subject of jack the ripper would be like saying michael jackson is starting to look a little odd . the book (or " graphic novel , " if you will) is over 500 pages long and includes nearly 30 more that consist of nothing but footnotes . in other words , don't dismiss this film because of its source . if you can get past the whole comic book thing , you might find another stumbling block in from hell's directors , albert and allen hughes . getting the hughes brothers to direct this seems almost as ludicrous as casting carrot top in , well , anything , but riddle me this : who better to direct a film that's set in the ghetto and features really violent street crime than the mad geniuses behind menace ii society ? the ghetto in question is , of course , whitechapel in 1888 london's east end .

Sentiment: POS

Supervised Learning (cont)

- Data Annotated/labeled with information of interest
  - Recidivism Risk Scores

LastName	FirstName	Sex_Code_Text	DisplayText	RawScore
Fisher	Kevin	Male	Risk of Violence	-2.08
Fisher	Kevin	Male	Risk of Recidivism	-1.06
Fisher	Kevin	Male	Risk of Failure to Appear	15
KENDALL	KEVIN	Male	Risk of Violence	-2.84

- Data used for training COMPAS system
  - Decision support tool in certain US Courts

Supervised Learning (cont)

- SL in a nutshell
  - Feed millions of annotated/labelled data examples to machine (algorithm)
  - Ask algorithm to learn which variables are most predictive
    - Words, parts-of-speech → sentiment
    - Age, education level, sex → recidivism risk
  - Process known as **TRAINING**
  - Various SL algorithms

Supervised Learning (cont)

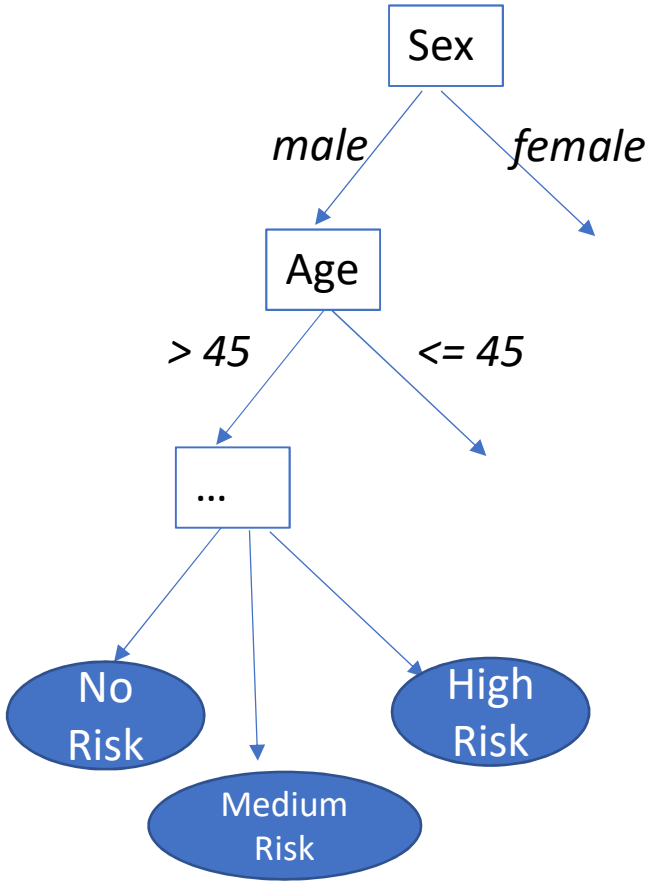
- Algorithms for Training
  - Tree-based
    - Decision-trees, Random Forest
  - Support Vector Machines
  - Neural Networks
  - And many others…

Tree-Based

- Learns a Decision-tree or Random Forest from training data

- Given new case
  - Follow tree branches/values
  - Make prediction

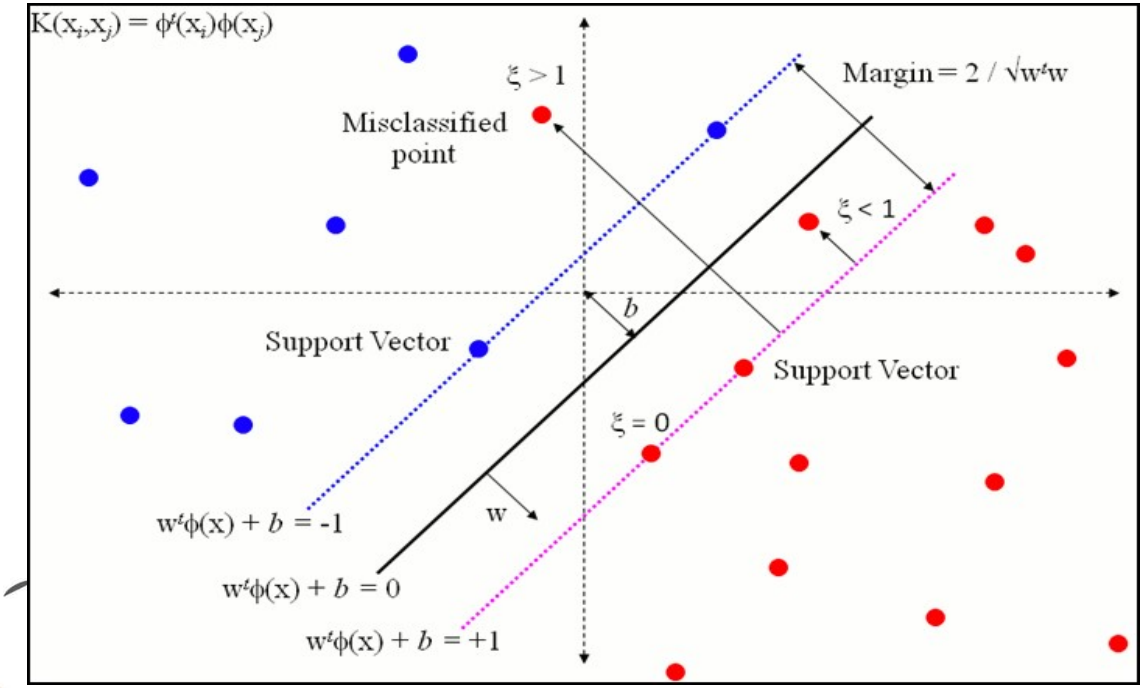
- Many variants
  - Gradient boosted trees
  - Ensemble of trees



Support Vector Machines (SVM)

- Learns decision boundaries

- Maximizes separation between examples
  - High vs. low risk
  - POS vs. NEG sentiment

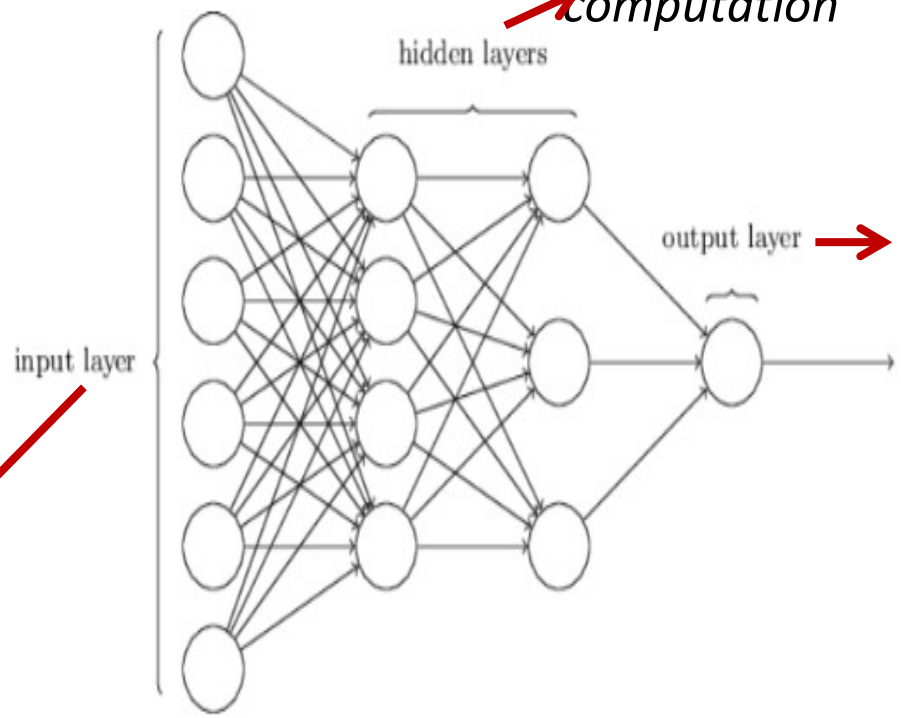


Source: stackexchange

Artificial Neural Networks

- Interconnected neurons

Responsible for most of the computation

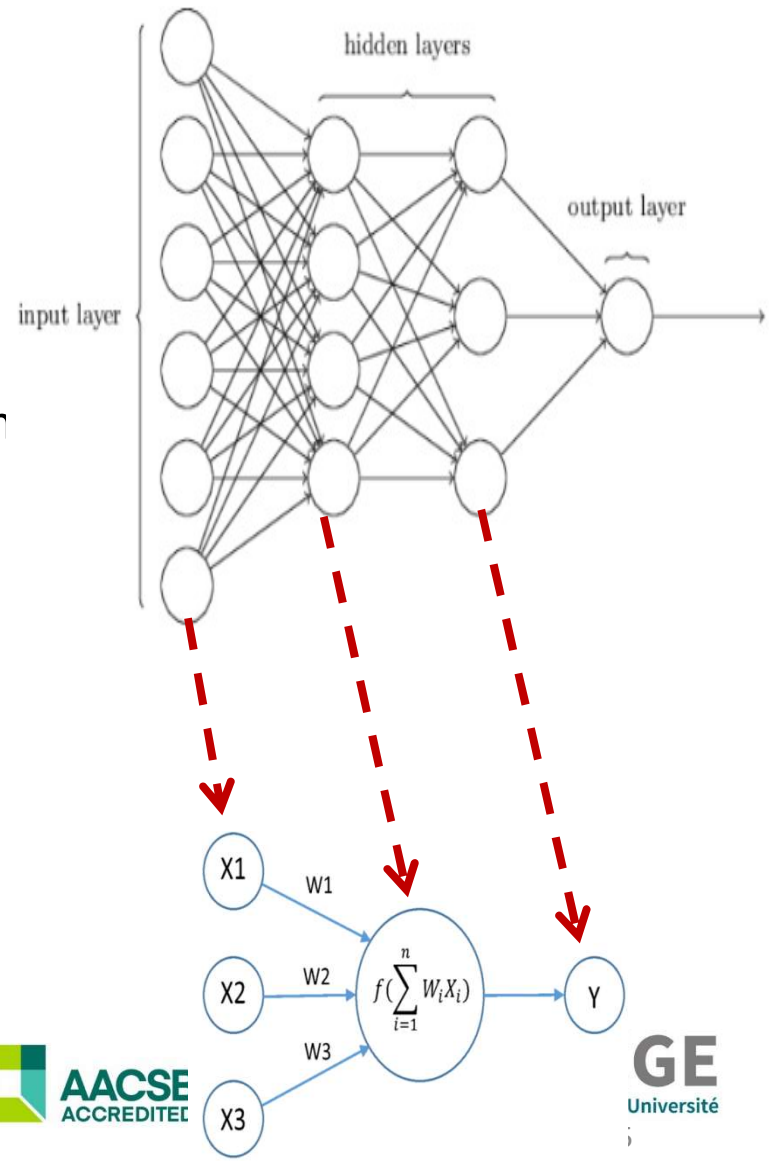


Tries to predict dependent variable, e.g. recidivism risk

Independent variable, e.g. age, past cases, sex,..

Artificial Neural Networks (cont)

- At each neuron *j*
  - Outputs of previous neurons; *x1,x2, x3*
  - Apply *weights*; *w1,w2, w3*
  - Sum & Apply *activation function, f(.)*
  - Result input to next neuron
  - Repeat until final output neuron (prediction
- At output layer, check predicted value
  - Dependent variable value from dataset
- If prediction ok: stop
- Else: back-propagate error
  - Adjust weights
  - Until correct prediction



Deep Learning

- Artificial Neural Networks
- Large number of hidden layers (e.g. 10, 300)

- Performs more sophisticated/complex tasks
  - Outperformed human champion at game of GO
  - Learns masters level chess by *playing against itself*

- However
  - High computational time & complexity
  - Difficult to interpret models

- Common Applications
  - Facebook Face Recognition
  - Voice/Speech recognition, e.g. Siri, Alexa

Supervised Learning Limitations

- SL matured learning paradigm

- But
  - Presupposes annotated data
  - Data in real-life not annotated
  - Manual annotation expensive, time-consuming

- Unsuitability of SL for many applications
- Need for other learning paradigms

UNSUPERVISED/SELF-SUPERVISED LEARNING

Unsupervised/Self-Supervised

- No specific unsupervised methods
  - Clustering
  - Similarity metrics (cosine, Jacquard)
  - PCA
  - LDA
  - Energy-based methods
  - Auto-encoders
  - …

- Aim
  - Not so much on prediction
  - But discovering patterns in data

Unsupervised/Self-Supervised

- Self-Supervised Learning
  - Unsupervised learning variant
  - Data provides the supervision

- Finding semantically-similar words
  - Synonyms
  - Near-synonyms

Unsupervised/Self-Supervised

- Distributional Semantics
  - "Similar words tend to occur in similar contexts"

- ***"In the new regulation*** assets are defined as properties…"
- "***The regulation*** refers to derivatives as products…"
- "Counterparty is specified ***in the regulation*** as …"

- Target words *defined (as), refer (to), specified (as)*
  - Share similar meaning
  - Share similar ***contexts***

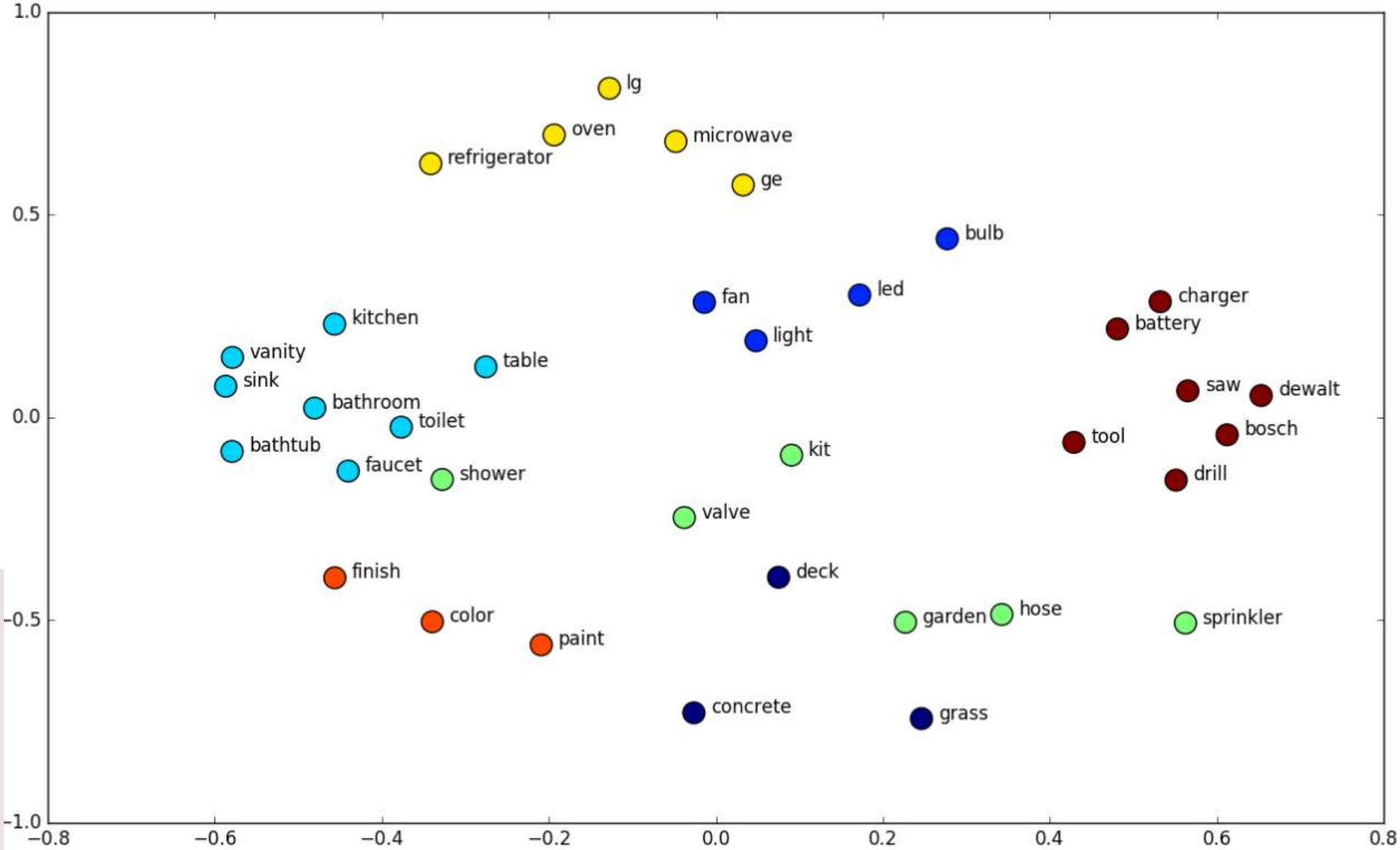
- Supervised Learning Formulation
  - Target words = annotations
  - Predicted from contexts

Unsupervised/Self-Supervised

- Train a neural network
  - Predict target words given contexts
- Target words represented as word-embeddings
  - Low-dimensional vectors, capturing semantics
- Well-known methods for generating word-embeddings
  - Word2Vec (Google)
  - FastText (Facebook)
  - GLOVE (Stanford U)
- *Methods used in Eur-Lex Project*

Word Embeddings Plot

- Vectors (embeddings) of similar words are close to each other



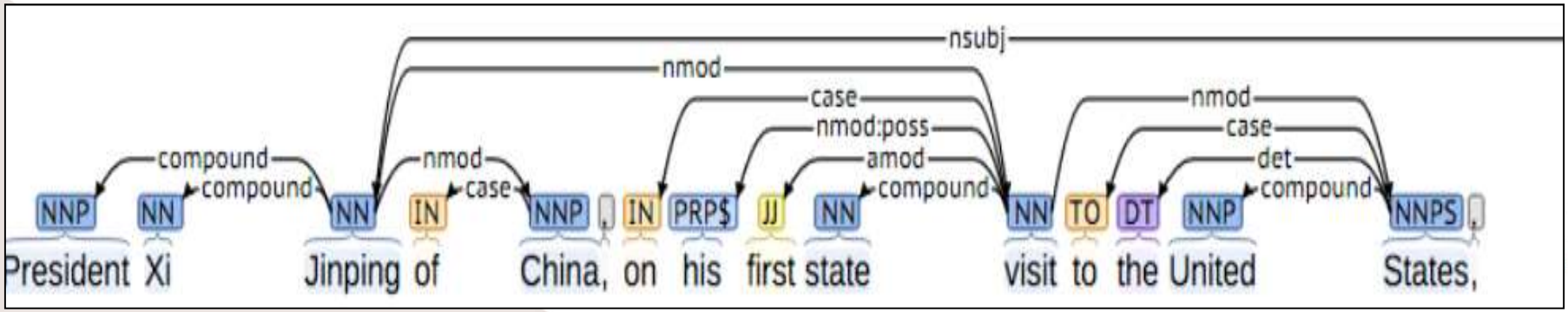
Natural Language Processing (NLP)

- Processing natural language, texts
- Machine Learning & Deep Learning
- Supervised & unsupervised

- Several sub-tasks
  - Part-of-Speech (POS) Tagging

```
Regulation/NNP shall/MD not/RB affect/VB the/DT accounting/NN standards/NNS applied/VBN by/IN supervised/JJ entities/NNS
```

  - Syntactic Parsing



End of Part 1

Questions?

Part 2: Eur-Lex Machine Learning Project

Project Team

- Myself
- PhD and Masters researchers at ULiège

EUR-Lex ML Project Overview

- Objective
  - Automatically analyze given set of EUR-Lex financial sector regulations
- Identify relevant legal concepts
  - Related to supervisory reporting requirements
- Organize concepts in dictionary
  - With references to “where” they occur
  - URL, Articles or Section numbers
- Extract concept definitions
- Extracting reporting information
  - “banks shall submit their filings to the central authority”
- Overarching Aim
  - Generate supervisory reporting concept dictionary

Methodological Overview

- Main methods employed
  - NLP
  - Unsupervised & self-supervised setting
  - No annotations available
  - Manual annotations not viable (expensive, time-consuming)

- Implemented in an NLP pipeline
- 3 main Work Packages (WP)/steps

Work Packages

- Data Gathering
- Concept Extraction
- Relation extraction
  - Definitions
  - Report-to

WP1: Data Gathering

- Eur-Lex regulations published online (HTML)
  - Example

9.11.2018 EN Official Journal of the European Union L 281/1

COMMISSION IMPLEMENTING REGULATION (EU) 2018/1627

of 9 October 2018

amending Implementing Regulation (EU) No 680/2014 as regards prudent valuation for supervisory reporting

(Text with EEA relevance)

THE EUROPEAN COMMISSION,

Having regard to the Treaty on the Functioning of the European Union,

Having regard to Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012 ([1]) and in particular the fourth subparagraph of Article 99(5), the fourth subparagraph of Article 99(6), the third subparagraph of Article 394(4) the fourth subparagraph of Article 415(3) and the third subparagraph of Article 430(2) thereof,

Whereas:

Data Gathering (cont)

- Retrieve online HTML documents
- Store documents in local repository

- Implemented a web-crawler (spider)
  - Starts from a predefined URL list (621 for prototype)
  - Automatically selects EN version

  - Reads and parses HTML documents

- Enriches contents with additional meta-data
  - Useful for subsequent WP

Data Gathering (cont)

- Enriched HTML contents dumped locally
- Plain Text files (.txt)
  - Lightweight
  - Portable

- Crawled contents in text files

```
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=celex:32018R1627&from=EN/
9.11.2018
EN
Official Journal of the European Union
L 281/1
COMMISSION IMPLEMENTING REGULATION (EU) 2018/1627
of 9 October 2018
amending Implementing Regulation (EU) No 680/2014 as regards prudent valuation for su
(Text with EEA relevance)
THE EUROPEAN COMMISSION,
Having regard to the Treaty on the Functioning of the European Union,
```

Data Gathering (cont)

- Crawler Characteristics
  - Python implementation
  - Lightweight
  - Fast (621 documents approx. 1hr-1.5hr)
  - Parses PDF & HTML

WP2: Concept Extraction

- Identifying & extracting ***relevant legal concepts*** from crawled contents
- Concepts linguistically realized as ***terms***
  - Single-word: *derivative*
  - Multi-word: *credit derivative volume*

- NLP algorithms for Term Extraction
- Initial approaches
  - Linguistic
  - Statistical
  - (Linguistic & Statistical)

Concept Extraction (cont)

- Linguistic Approach
  - Terms are sequences of adjectives followed by any number of words
  - Determine Part-of-Speech of each word in documents

```
In|IN the|DT |NN of|IN futures|NNS and|CC forwards|NNS other|JJ than|IN futures|future|NNS

In|IN the|DT case|NN of|IN swaps|NNS related|VBN to|TO securities||NNS ,the|DT counterparty|NN
```

- Terms dectected
  - *Futures and forwads*
  - *Futures*
  - *Swaps*
  - *Securities*
  - *Counterparty (side)*

Concept Extraction (cont)

- Statistical Approach
- Terms detected according to
  - Occurrence frequencies /probabilities (single-word terms)
  - Co-occurrence frequencies/probabilities (multi-word terms)

- Numerous statistical measures
  - Mutual Information
  - Chi-square

- Good results for 2 or 3-word terms

```
technical standards
publication official
competent authorities
enter force
official journal
non-performing exposures
replaced text
remittance date
default impairment
```

Concept Extraction (cont)

- Several challenges posed by Eur-Lex project
- Terms not restricted to adjectives and nouns
- Very long terms
  - Composed of > 4 words
  - “Financial assets designated at fair value through profit or loss”
- Valid terms with low frequency/probability
- Annotated data/examples unavailable
  - Require unsupervised approaches

Concept Extraction (cont)

- Implemented novel, unsupervised Term Extraction algorithm

- Dynamic frequency threshold, t
  - Varies depending on the document length
  - Terms with frequency < t are discarded

- Keep tracks of terms' positions in document
  - Article, section, annex numbers
  - Character positions (in progress)

Concept Extraction (cont)

- Example Input to Algorithm (crawled HTML in txt format)

Official Journal of the European Union
L 233/1
COMMISSION DELEGATED REGULATION (EU) 2016/1434
of 14 December 2015
correcting Delegated Regulation (EU) 2015/63 supplementing Directive 2014/59/EU of the European Parliament and of the Council with regard to ex ante contributions to resolution fina
THE EUROPEAN COMMISSION,
Having regard to the Treaty on the Functioning of the European Union,
Having regard to Directive 2014/59/EU of the European Parliament and of the Council of 15 May 2014 establishing a framework for the recovery and resolution of credit institutions an
Whereas:
(1)
Some errors appear in all language versions of the text of Articles 5(1)(f), 5(3), 6(9), 12(1), 14(1), 20(1) and 20(5) of Commission Delegated Regulation (EU) 2015/63 (2).
(2)
Article 5(1)(f) of Delegated Regulation (EU) 2015/63 erroneously contains the word 'original' reducing thereby the scope of the exclusion relating to the liabilities of promotional
(3)
In Article 5(3) of Delegated Regulation (EU) 2015/63 the reference to Article 429(6) and (7) of Regulation (EU) No 575/2013 of the European Parliament and of the Council (3) should
(4)
In Article 14(1) of Delegated Regulation (EU) 2015/63, it should be clarified that it refers to the latest approved annual financial statements available, at the latest, on 31 Decem
(5)
Article 20(1) contains a typographical error. The deadline should be aligned with the deadline in paragraph 4 of that Article and changed to 1 September 2015.
(6)
Article 20(5) needs to be aligned with Article 8(5) of Council Implementing Regulation (EU) 2015/81 (5) in order to ensure consistency within the internal market and in Union law. T
(7)
Further errors appear in the German version of the text of Articles 14(1), 15(2) and 16(1) of Delegated Regulation (EU) 2015/63.
(8)
Delegated Regulation (EU) 2015/63 should therefore be corrected accordingly.
(9)
The errors in Delegated Regulation (EU) 2015/63 require a correction to ensure a level playing field in the internal market. For this reason, this Correcting Regulation should apply
HAS ADOPTED THIS REGULATION:
Article 1
Delegated Regulation (EU) 2015/63 is corrected as follows:
(1)
in Article 5(1), point (f) is replaced by the following:
'(f)
in the case of institutions operating promotional loans, the liabilities of the intermediary institution towards the originating or another promotional bank or another intermediary

Concept Extraction (cont)

```
18 | non-trading non-derivative financial assets measured at fair value
18 | official export credit insurance scheme
18 | relevant day of the reporting period with an original maturity
17 | total direct loss recovery
17 | recovery from insurance and other risk transfer mechanisms
```

```
grandfathered instruments
-----------------
annex i
annex ii

transitional provisions
-----------------
annex i
annex ii

risk-weighted exposure
-----------------
annex i
annex v
annex ii
```

Concept Extraction (cont)

- Algorithm's Characteristics
  - Implemented in Python
  - Lightweight, easy to install
  - Efficient
    - Uses memoization
    - ~35 000 characters analyzed per second

WP3: Semantic Relationship Extraction

- ~~Data gathering~~
- ~~Identifying concepts (terms) & positions~~
- Extracting
  - Concepts' definitions
  - Reporting information *(concept A shall submit to concept B…")*

- Semantic Relationship Extraction in NLP

- No annotated data for current project
- Devise unsupervised methods

Semantic Relationship Extraction

- Extracting Concepts' Definitions
- Various lexical patterns to express definitions
  - "In the new regulation assets ***is defined as*** properties…"
  - "The regulation ***refer to*** derivatives as products…"
  - "Counterparty ***is specified*** in the regulation ***as*** …"

- Challenge:
  - Learn how definitions are expressed in documents automatically
  - Unsupervised fashion

- Solution: Word Embeddings

Extracting Definitions

- Word embedding
  - Low dimensional vector representation of words
  - Vectors inherently captures semantic information
  - Generated via neural networks
    - E.g. Predict word given its contexts

Extracting Definitions (cont)

- Investigated different methods
  - Word2Vec (Google)
  - FastText (Facebook)
  - GLoVE (U. Stanford)

- Generate word embeddings for words from
  - Google news
  - Wikipedia
  - WebCrawl
  - Eur-Lex corpus of financial sector regulations

Extracting Definitions (cont)

- Investigated different methods
  - Word2Vec (Google)
  - FastText (Facebook)
  - GLoVE (U. Stanford)

- Generate word embeddings for words from
  - Google news
  - Wikipedia
  - WebCrawl
  - Eur-Lex corpus of financial sector regulations

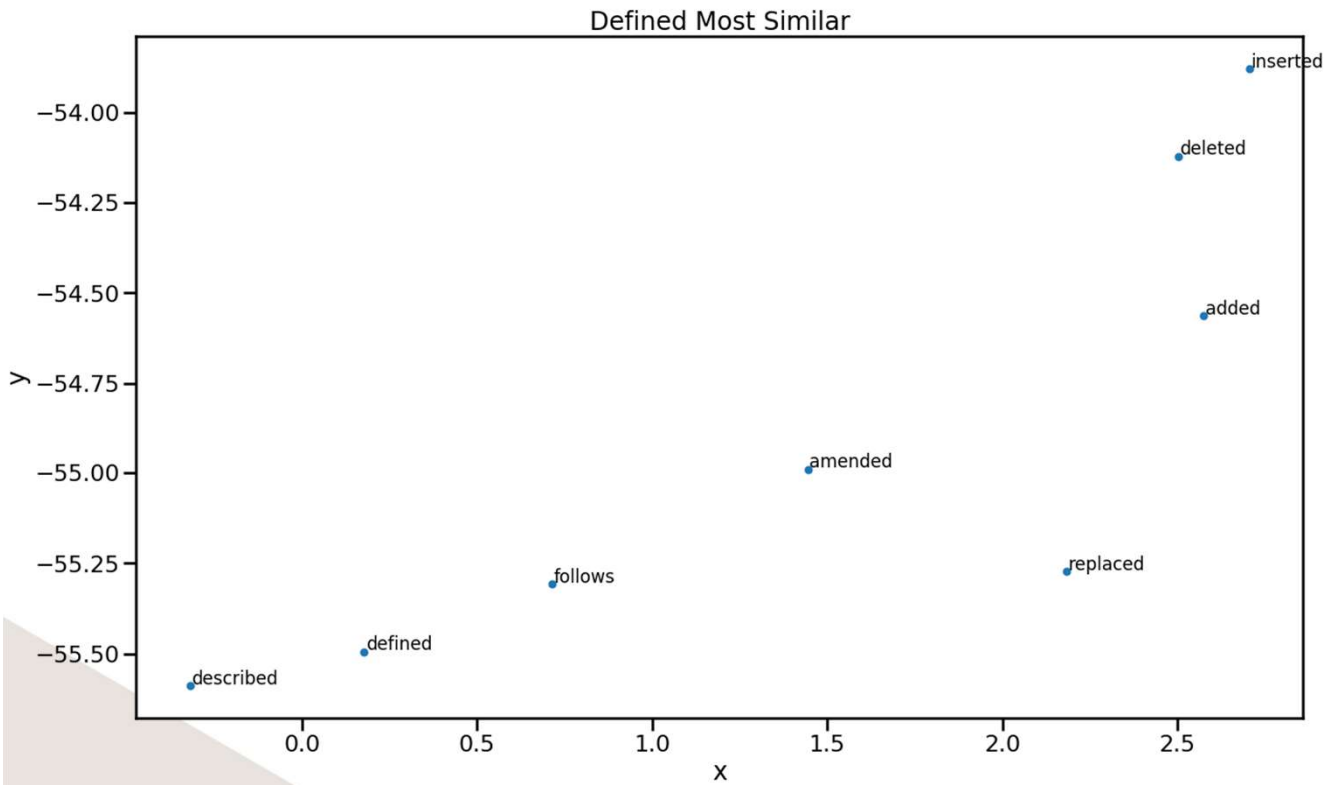
Extracting Definitions (cont)

- Explicit definition word: “(to) define”
- Look up its vector (word embedding), *v*

- Compute distance(v, x), x = embedding of all other words

- If distance(v,x) < threshold
  - x: vector word w
  - Similar meaning as “define”

- Distance: Cosine of the angle between vectors



defines
define
defined
defining
characterized
categorized
redefined
specified
interpreted
governed
denoted
understood
called
dubbed
equated
regarded
known
termed
labeled
means

Extracting Definitions (cont)

- Recap
  - Concepts extracted
  - Identified how “definitions” are expressed
  - Straightforward identification of concepts definitions in corpus

"Securitisation" means securitisation as defined in Article 4(1)(61) of Regulation

"Contractual netting agreements" means contractual netting agreements as defined in defined in Article 2(a) of Directive 98/26/EC or other payment or securities

The counterparties of the institution may agree on a so-called 'catch-all' or sweep-up securitisations by ***means*** of contractual instruments, where the underlying

False Positive
→ Need only verbs

Extracting Reporting Information

- Same procedure as Definition Extraction

Evaluation of Results

- Automatic
- But requires domain-knowledge for objective evaluation

End of Part 2

Questions?